Quantifying Culture

(Class 2)

Looking at data - numerical summaries and graphs

Alessio Palmisano
University of Turin

UNIVERSITÀ DI TORINO

Tech T4 Culture

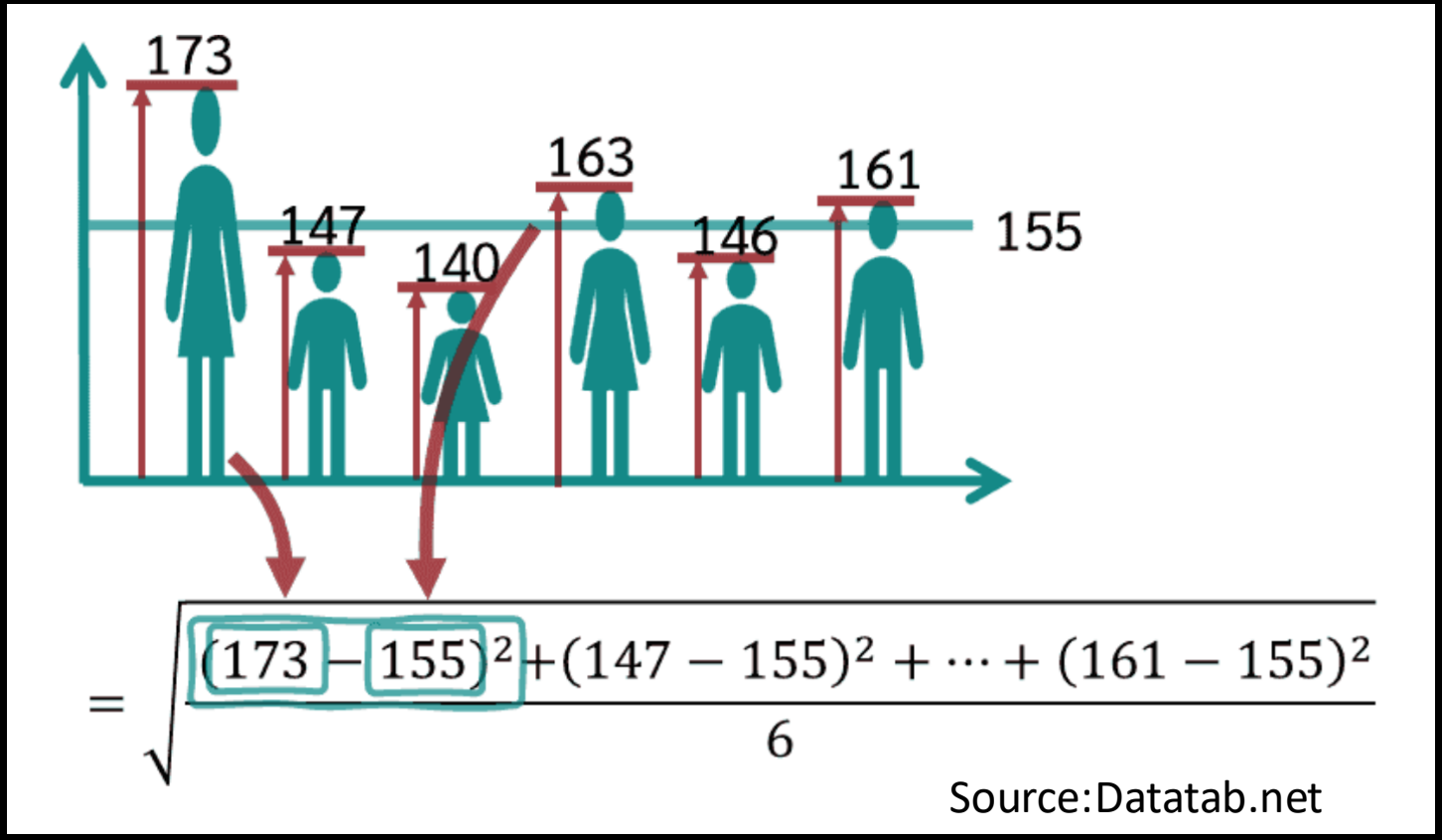
Outline

- Descriptive statistics
  - Standard deviation
  - Boxplot
  - Frequency distributions
- Binomial Distribution
- Poisson Distribution
- Sampling techniques

Standard deviation

The standard deviation tells us how much the data varies from the mean (average).

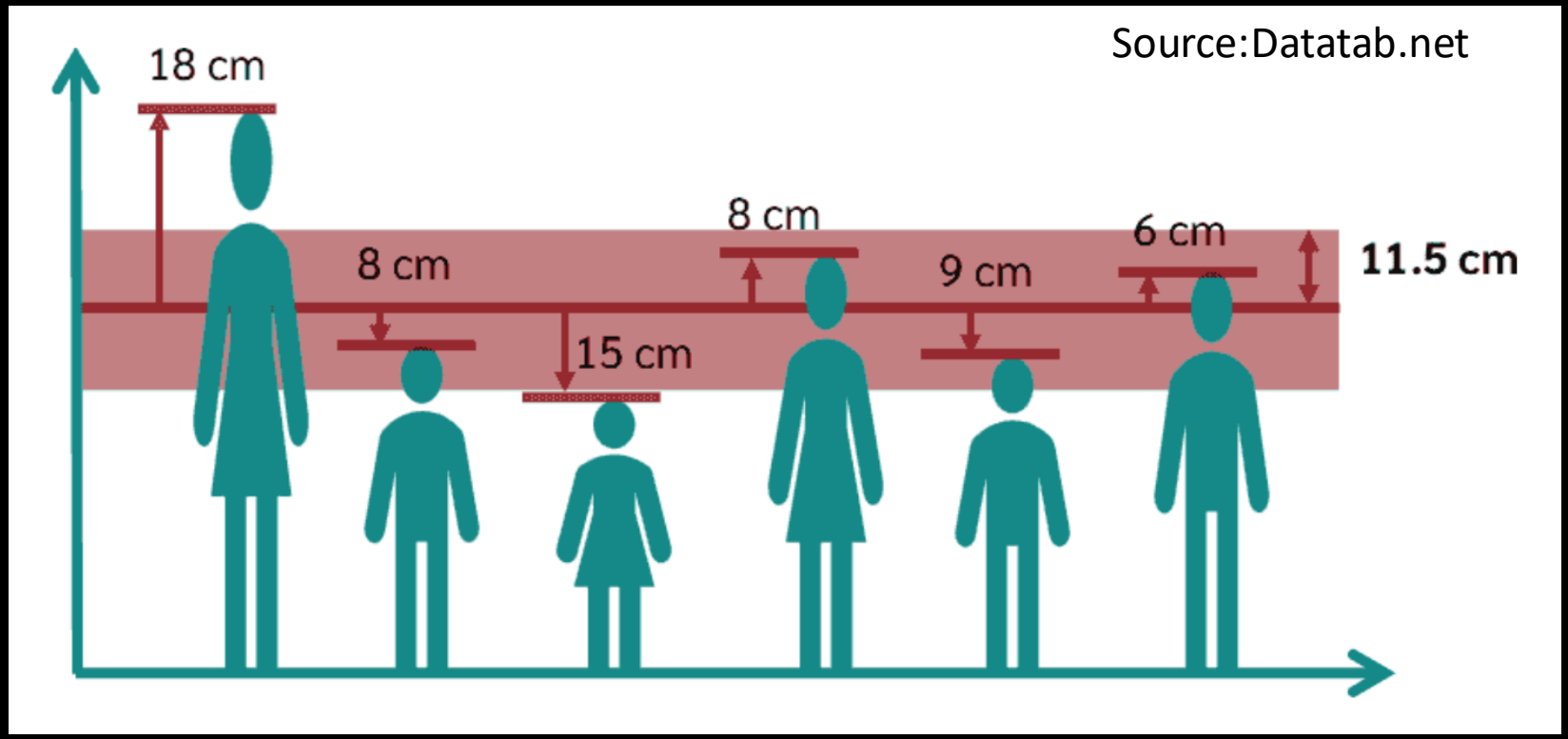
It reflects how far, on average, each individual value deviates from the mean, offering insight into the spread of the data.



Standard deviation

The standard deviation tells us how much the data varies from the mean (average).

It reflects how far, on average, each individual value deviates from the mean, offering insight into the spread of the data.



Boxplot

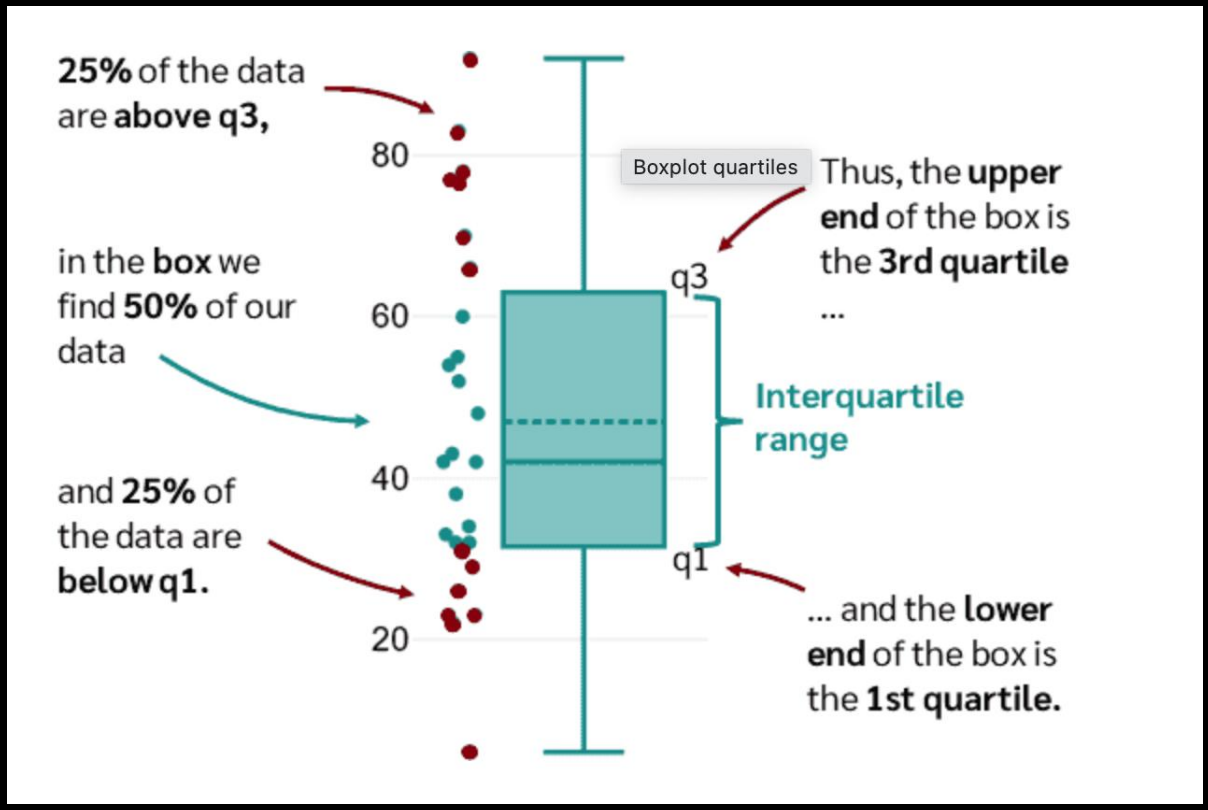
With a boxplot you can graphically display, the median, the interquartile range (IQR) and the outliers.

The **median** is the middle value of an ordered dataset. It divides the data into two equal halves.

- **Q1 (First Quartile):** 25th percentile (median of the lower half).
- **Q3 (Third Quartile):** 75th percentile (median of the upper half).

Applications of IQR & Box Plots

- Detecting outliers
- Comparing distributions in different datasets.
- Summarizing large data

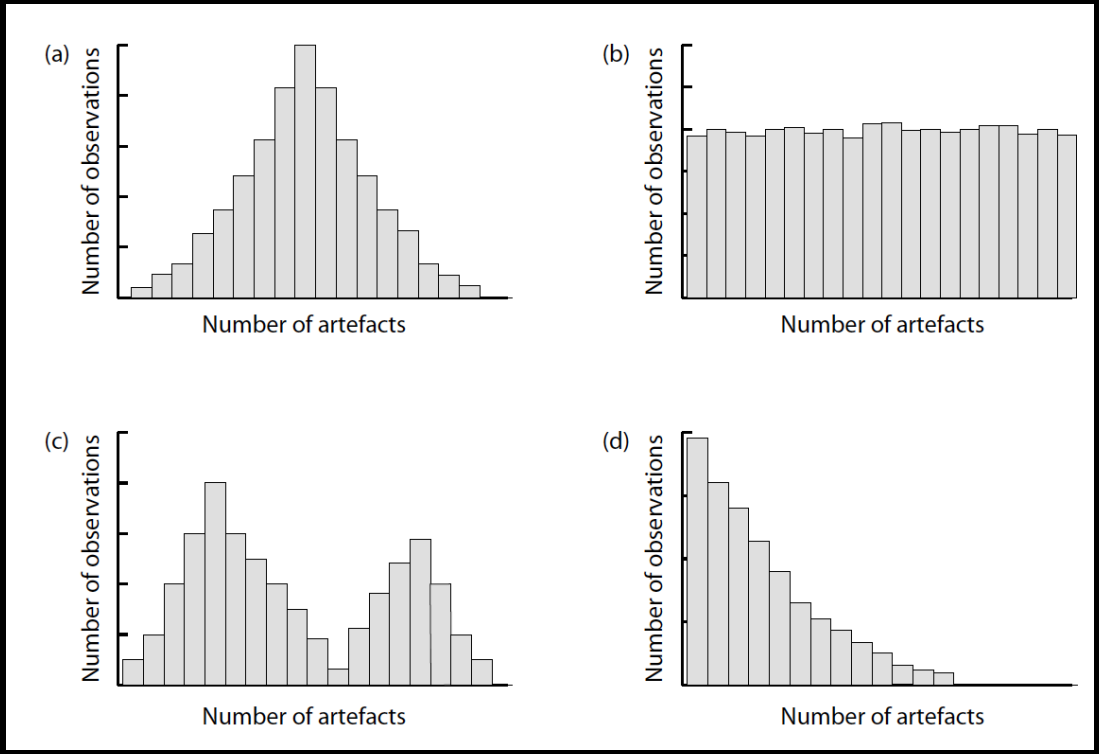


Coding time

- Tutorial pp. 2-4

Numerical Data - frequency distributions

Normal



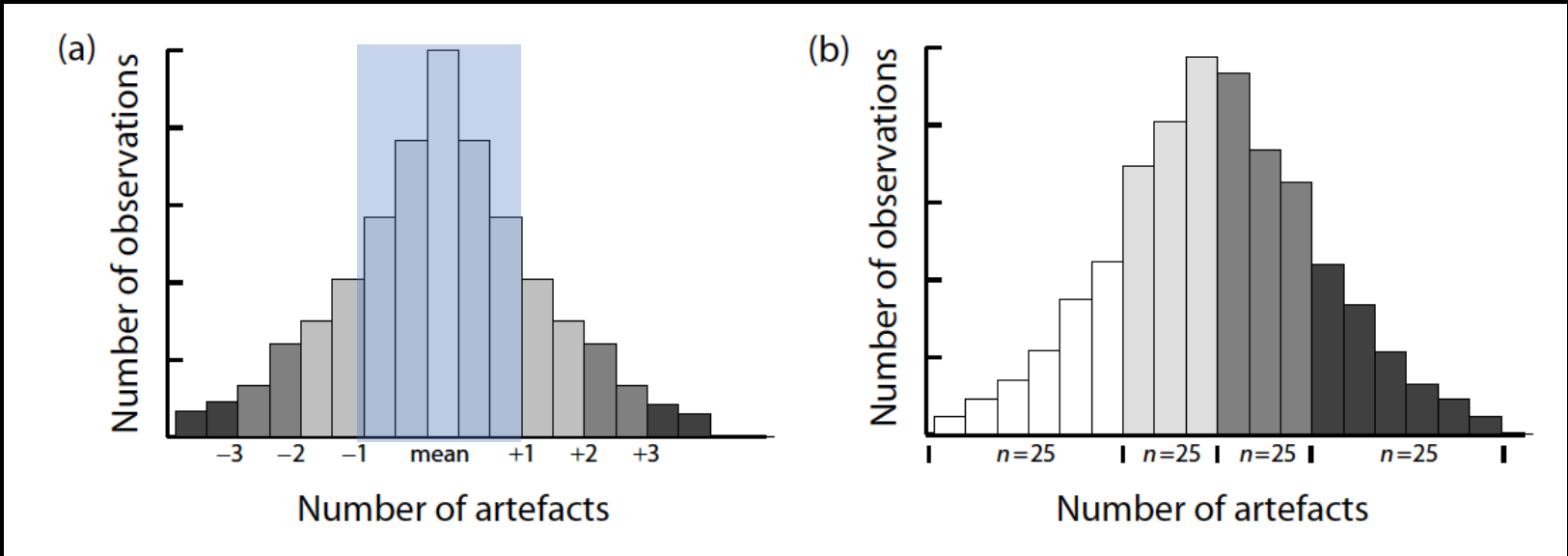
Rectanguar

Bimodal

Skewed

Frequency distributions and their effects on data generalisation

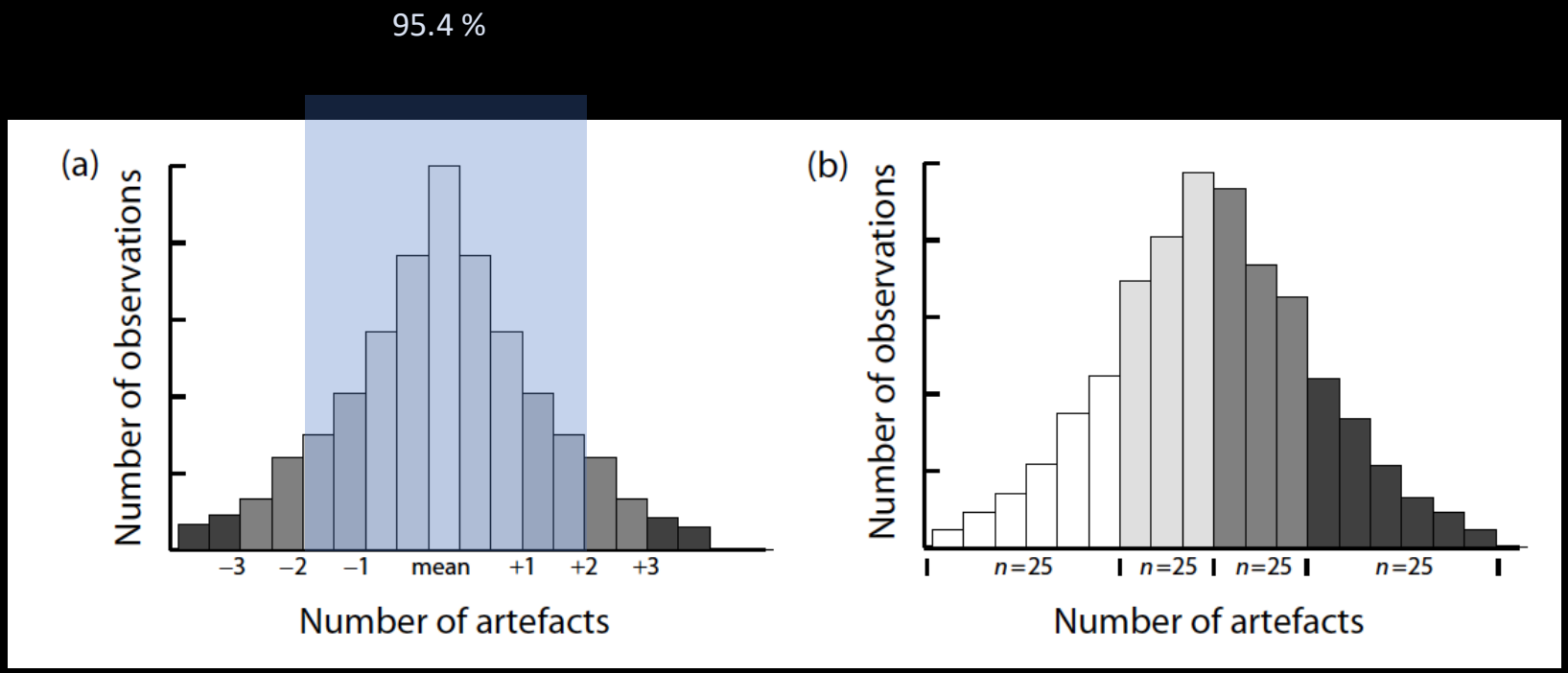
68.2 %



Standard deviation

Quartile

Frequency distributions and their effects on data generalisation



Standard deviation

Quartile

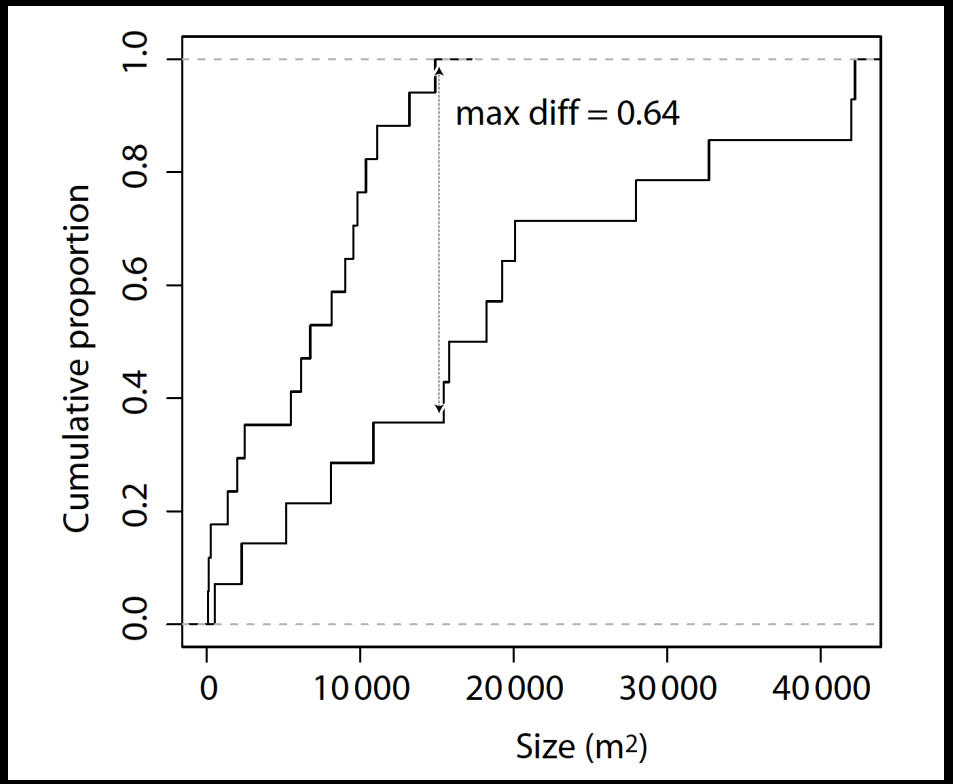
Coding time

- Tutorial pp. 4-9

Empirical cumulative frequency distribution (ECDF)

(ECDF) is a statistical function that represents the proportion or count of observations in a dataset that are less than or equal to a given value

- It starts at 0 and increases to 1 (or 100% if expressed as a percentage).
- It increases in **steps** at each data point, with the step height corresponding to the frequency of the observation.
- Directly shows how data is **distributed**.
- **More Informative** than histograms: No need for binning, providing a clearer view of **distribution shape**.



Kolgomorov-Smirnov (K-S) Test

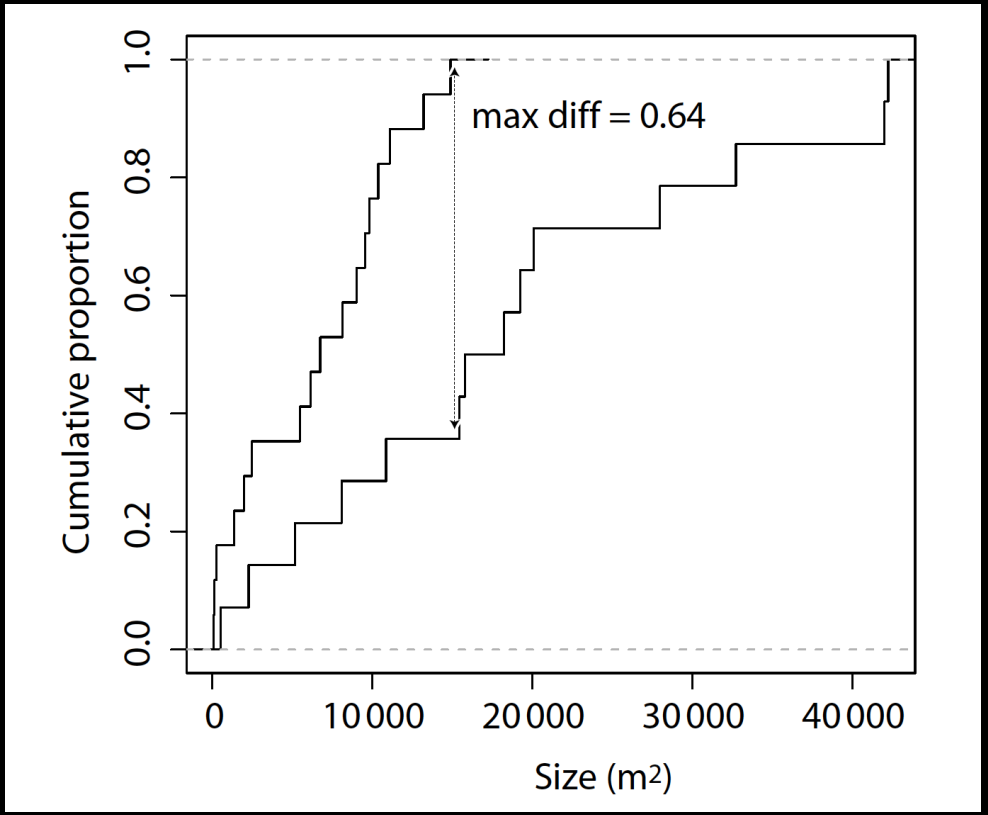
It is a non-parametric statistical test used to compare a sample distribution with a theoretical distribution (**one-sample K-S test**) or to compare two empirical distributions (**two-sample K-S test**).

- It measures the **maximum difference** between the empirical cumulative distribution functions (ECDFs)

$$D = \max|S_1(x) - S_2(x)|$$

D = maximum vertical difference between ECDFs
S_1 (x) and S_2 (x) are the ECDFs of the two samples

- **Non-parametric**: No assumption about the data distribution.
- It assesses if two samples come from the **same distribution**.



Coding time

- Tutorial pp. 10-13

Binomial Distribution

It is a discrete probability distribution that describes the **number of successes** in a fixed number of independent **trials**, where each trial has only two possible outcomes: success or failure.

- **Fixed number** of independent trials.
- **Independent Trials**: the outcome of one trial does not affect the outcome of another.
- **Two Possible Outcomes**: Each trial results in either success (with probability *p*) or failure (with probability 1−*p*).
- **Constant Probability**: The probability of success *p* remains the same for all trials.

The probability of getting exactly k successes in n trials is given by:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where:

- X is the number of successes.
- $\binom{n}{k}$ is the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- p is the probability of success.
- $(1-p)$ is the probability of failure.
- k is the number of successes ($0 \leq k \leq n$).

Binomial Distribution

It is a discrete probability distribution that describes the **number of successes** in a fixed number of independent **trials**, where each trial has only two possible outcomes: success or failure.

Suppose you flip a fair coin ($p = 0.5$) **10 times** ($n = 10$), and you want to find the probability of getting **exactly 6 heads**.

Using the formula:

$$P(X = 6) = \binom{10}{6}(0.5)^6(0.5)^4$$

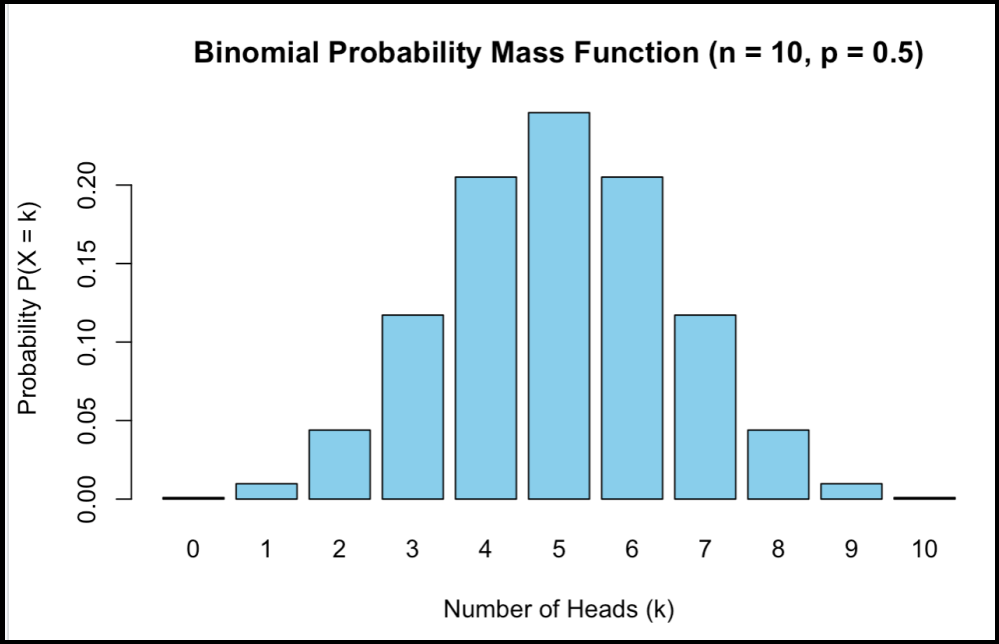
$$= \frac{10!}{6!(4!)}(0.5)^{10}$$

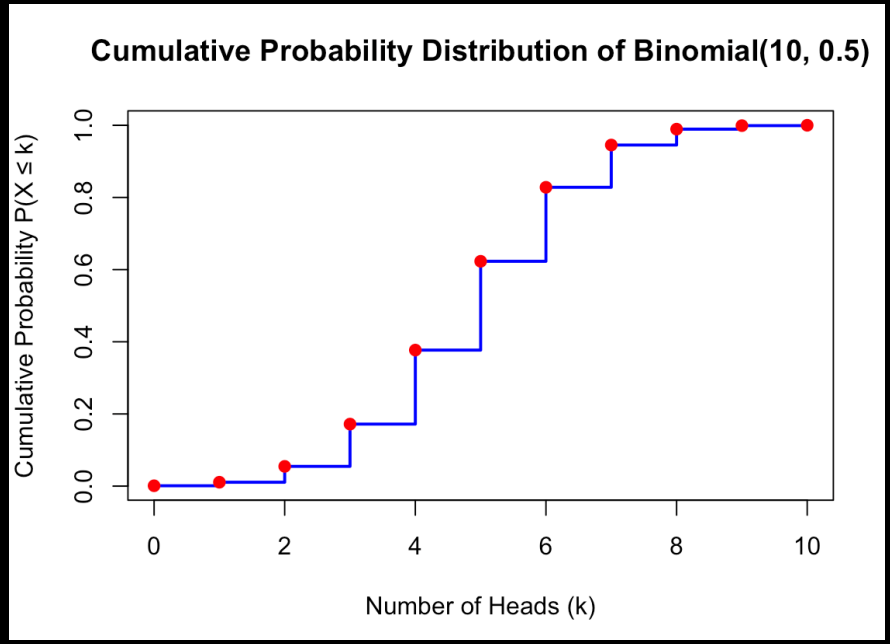
$$= 210 \times 0.000976 = 0.205$$

So, the probability of getting exactly 6 heads is **0.205** (or **20.5%**).

Binomial Distribution

It is a discrete probability distribution that describes the **number of successes** in a fixed number of independent **trials**, where each trial has only two possible outcomes: success or failure.





Coding time

- Tutorial pp. 18-21

Poisson Distribution

The **Poisson distribution** is a discrete probability distribution that models the **number of events** occurring in a **fixed interval of time or space**, given that the events occur independently and at a **constant average rate**.

- To model the **random distribution of artifacts** across space and time.
- To determine whether pottery fragments are **evenly distributed or clustered** within an excavation site.
- The occurrence of fossil or human remains over time
- The probability of finding settlements within a given square kilometer.

The probability of observing exactly k events in a given interval is:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

where:

- X is the number of occurrences,
- λ is the average number of occurrences in the interval,
- e is Euler's number (≈ 2.718),
- $k!$ is the factorial of k.

Poisson Distribution

The **Poisson distribution** is a discrete probability distribution that models the **number of events** occurring in a **fixed interval of time or space**, given that the events occur independently and at a **constant average rate**.

If archaeologists usually find 5 artifacts per 10m², the Poisson model can estimate the probability of discovering exactly 7 artifacts in another 10m² area using:

- $\lambda = 5$ (average number of artifacts per 10m²)
- $k = 7$ (the exact number of artifacts we want to find)
- The formula for the Poisson probability is:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Now let's calculate it:

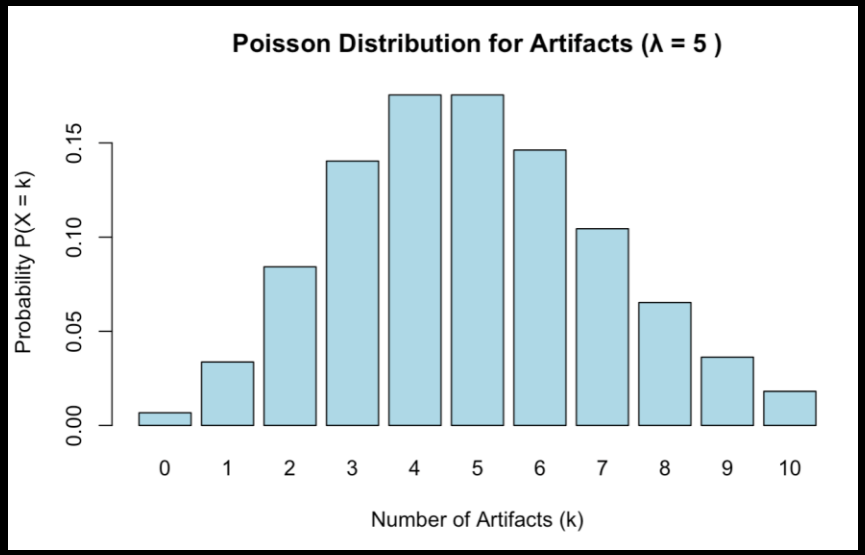
$$P(X = 7) = \frac{e^{-5}5^7}{7!}$$

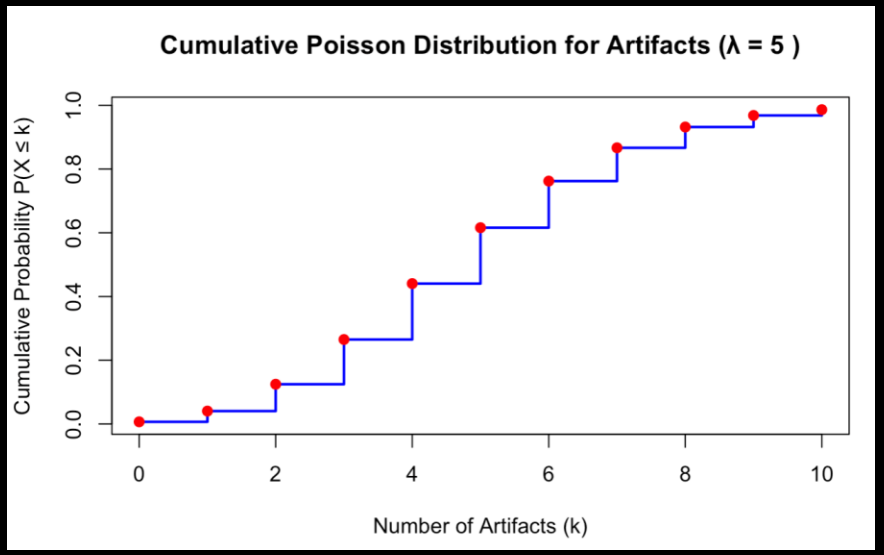
$$P(X = 7) = \frac{0.0067 \times 78125}{5040} \approx 0.101$$

Poisson Distribution

The **Poisson distribution** is a discrete probability distribution that models the **number of events** occurring in a **fixed interval of time or space**, given that the events occur independently and at a **constant average rate**.

If archaeologists usually find 5 artifacts per $10m^2$, the Poisson model can estimate the probability of discovering exactly 7 artifacts in another $10m^2$ area using:





Coding time

- Tutorial pp. 21-24

Why do we sample?

What are the stereotypical attitudes, amongst archaeology, about why we sample?

- • a black art …
- • an alien imposition..
- • a regrettable necessity…
- • a passport to science
- • an escape from tedium

Orton 2000: 4-5

Orton, C., 2000. *Sampling in Archaeology*. Cambridge: Cambridge University Press.

Why do we sample?

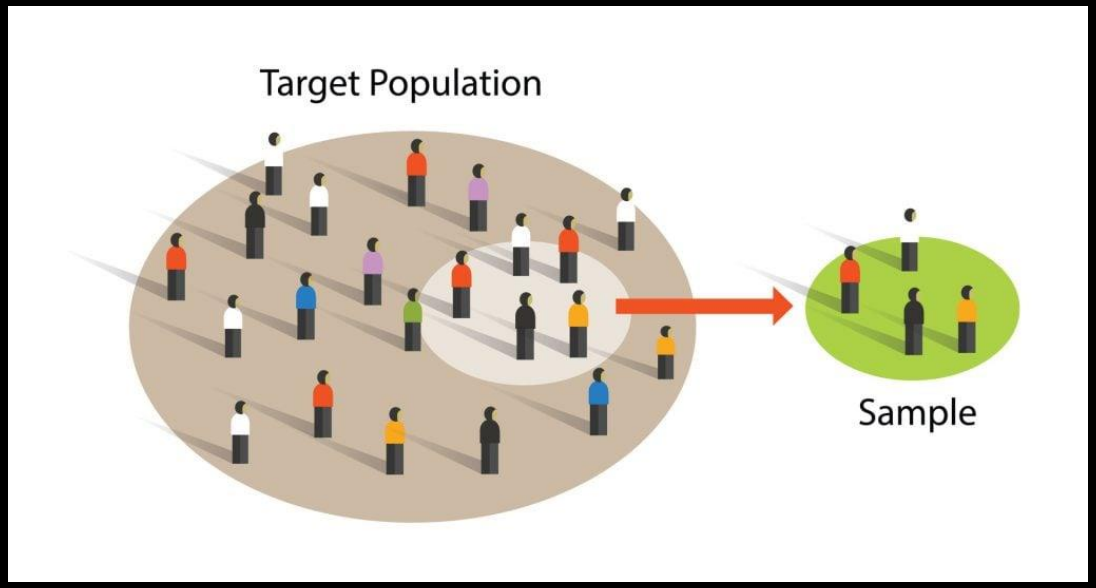
We do not have ‘everything’.

- ’total’ survey,
- ‘total’ excavation,
- ‘total’ artefact to analyse
- ‘total’ retrieval of objects from features,
- scientific analyses of ‘every’ item, etc.

Why do we sample?

Sampling is the process of **selecting** a **subset** of individuals, items, or observations from a larger population to study and **draw conclusions** about the **whole group**.

Since studying an entire population can be time-consuming, costly, or impractical, sampling provides a more efficient way to analyse data while still obtaining reliable results.



What makes a 'good' sample when we cannot examine everything?

- Do some samples offer more value than others?
- If they do, what does that suggest about their quality?
- How can we determine which sample is more suitable?
- Is the quality determined by the sample itself, or by the method used to select it?
- This idea is often framed as the need for our samples to be 'representative.' But what does 'representative' actually mean?
- What methods can we use to ensure our sample is representative? And how can we be certain of its accuracy?

Same fundamental concepts

- **Population**: The entire set of 'objects' or entities that are the focus of our research, from which we seek to gather information.
  - **target population** (the group we aim to learn about)
  - **sampled population** (the group we actually sample due to practical limitations).

- **Sample**: the outcome of one selection procedure, which consists of a set of units and the values associated with them.

- **Sample design:** defining the amount of units to be selected as well as the method of selection.

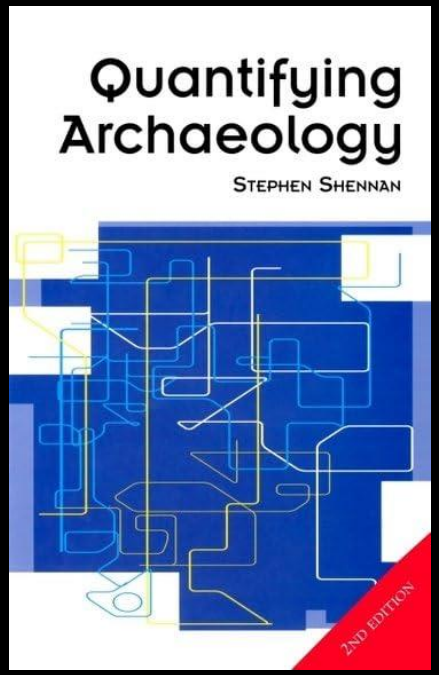
Sampling strategy

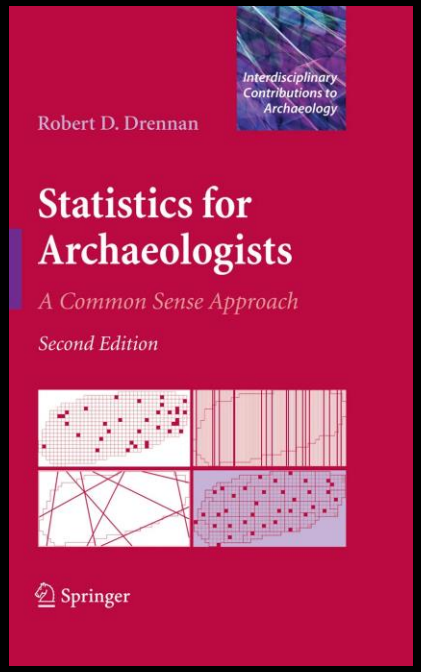
- **Simple Random Sample**: A sampling method where each unit has an equal chance of being selected.
- **Systematic Sample**: A method where units are selected at regular intervals across the sampling frame.
- **Stratified Sampling**: It involves dividing the population into distinct subgroups, known as strata, and sampling each separately. When the sampling fraction remains consistent across all strata, it is called "proportional allocation”.

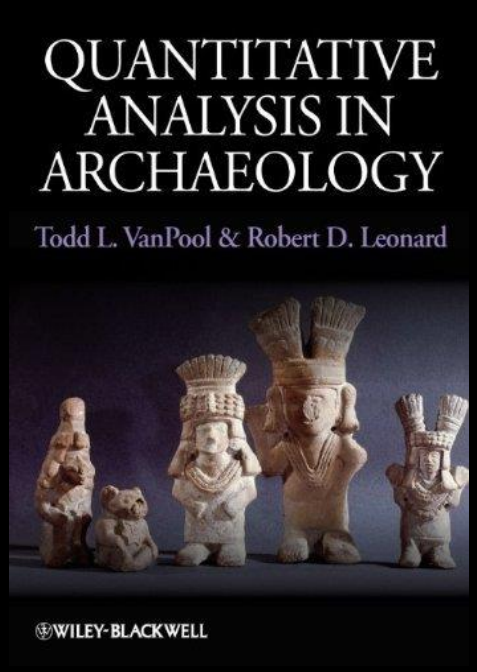
Coding time

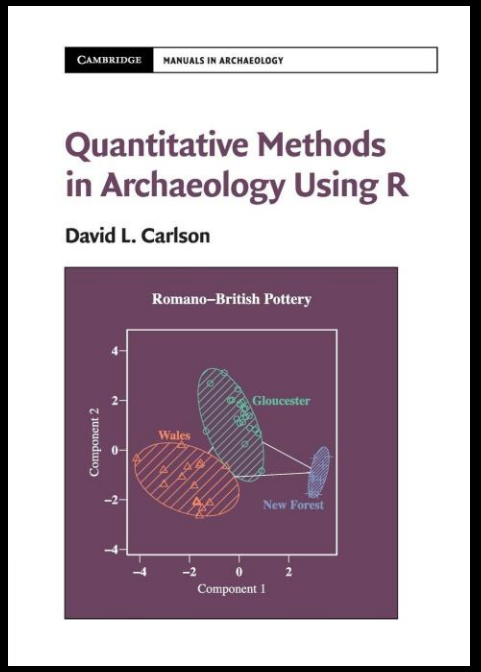
- Tutorial pp. 25-30

Statistical textbooks in Archaeology









- Shennan, S. (1997). *Quantifying archaeology*. University of Iowa Press.
- Drennan, R. D. (2010). *Statistics for archaeologists*. New York: Springer.
- Van Pool, T. L., & Leonard, R. D. (2011). *Quantitative analysis in archaeology*. John Wiley & Sons.
- Carlson, D. L. (2017). *Quantitative methods in archaeology using R*. Cambridge University Press.